Development of Unsupervised Classification Techniques for Water Quality Analysis

by Perry LaPotin[1] and Robert H. Kennedy[2]

PURPOSE: In large thematic studies, acquiring water quality data is costly and time-consumptive. For this reason, remote sensing is used as an initial estimation technique to identify significant environmental parameters. As shown in Cox et. al. (1998), correlation/regression may be used as a statistical tool to link the radiometric measures (acquired by the remote sensing platform) to the in situ measures (acquired by field sampling). This approach is both sensible and highly effective when an adequate sampling density is used that provides significant coverage for the study region. However, most studies cannot provide this coverage since the acquisition is far too costly and/or time-consumptive to meet the stochastic measures for sampling accuracy and precision. In addition, the related problem of how best to design a spatial sample within a highly diverse system (heterogeneous environmental strata) is an ongoing field of research (Bruce and Gao 1998).

This technical note examines an innovative method of performing unsupervised classification that requires no a priori field sampling. The approach is self-contained and is not statistically biased by a priori field samples. However, the technique is highly conducive to the use of an a posteriori sample of field data to refine the unsupervised classes. In this regard, the technique may be used to: (1) estimate coarse surface features that require detailed field sampling, and (2) locate unusual patterns from within large regions that require in-depth field sampling to identify cause-and-effect relationships.

The unsupervised method differs significantly from traditional methods applied within commercial image processing applications (Smith and Cole 1999). The approach does not use random sampling and is not based upon equal area sampling. In addition, no a priori assumptions are made concerning the normality (symmetry, kurtosis) of the underlying image data. Rather, a focused methodology is applied that looks for optimal separation of patterns and features.

BACKGROUND: Many reservoirs are large, morphologically complex ecosystems exhibiting pronounced spatial and temporal trends in limnological characteristics (Kennedy, Thornton, and Ford 1985; Søballe et al. 1992). Longitudinal gradients in water quality are common for long reservoirs receiving inflows from a single large tributary, especially when the point of inflow is distant from the dam. This is due primarily to the effects of sedimentation and density currents (Kennedy and Walker 1990), since both processes result in declines in nutrients and suspended material in surface waters along the axis of the reservoir. Reservoirs with two or more hydrodynamically distinct basins, such as embayments formed by the flooding of secondary tributaries, frequently exhibit great spatial heterogeneity due to the differential effects of material loading and reduced mixing (Kennedy, Thornton, and Ford 1985).

[1] United Imaging Corporation, Lyme, NH.

[2] U.S. Army Engineer Research and Development Center (ERDC), Environmental Laboratory, Vicksburg, MS.

In traditional limnological investigations, one or more sampling stations are established at locations representing homogeneous regions within the reservoir. However, the number of stations established and subsequently sampled is often limited by logistics or funding, and spatial patterns are frequently poorly represented by the samples collected. In addition, reservoirs are dynamic systems in which water quality patterns are strongly influenced by hydrologic and operational events. As a result, patterns in the distribution of water quality variables in most reservoirs exhibit significant temporal change. Thus, traditional limnological investigative methods may not adequately address concerns related to spatial and temporal trends in water quality.

Reflectance information, acquired using either satellite or airborne sensors, provides an alternative method for describing spatial patterns in reservoir water quality. LANDSAT Thematic Mapper (TM) provides information for seven wavelength bands at 30-m resolution. TM data, especially those for bands 1-4 and 6, have been used to describe patterns in the distribution of water temperature, turbidity (Feeney 1992), and algal pigment (Cox et al. 1998, Kennedy et al. 1994) in surface waters. Such applications have generally involved the development of empirical models relating observed water quality values and reflectance values for one or more wavelength bands. However, these models often have limited application when applied to image data for dates other than that upon which the model was developed (e.g., Kennedy et al. (1994)). This limitation reduces the potential application of image data, since coincident water quality data would be required, and cost savings would be minimal.

METHODOLOGY: Classification techniques are commonly applied to remotely sensed data to synthesize and reduce large volumes of information into a compact useable form. For multispectral and hyperspectral data, the technique is an essential step for reducing radiometric data into a format that is conducive to environmental analysis.

Water quality data may be merged with remotely sensed data using two primary analysis techniques: *supervised* classification and *unsupervised* classification. In the former case, it is assumed that in situ ground samples are available and of sufficient quality (precision) that the remotely sensed data should be classified (re-mapped) based upon the relevance of this information. In the latter case, the analyst assumes that the in situ data are not available, of poor precision, or are too costly to obtain within the required time interval (at the prescribed sampling scale). For these conditions, the analysis is based purely upon the accuracy and precision of the image data. For the unsupervised case, the imaging system is used as a sole-source method for the identification of patterns and features.

Supervised Classification. The supervised approach assumes that the fidelity of the in situ measurement is superior to the free-weight classification of the image data. Restated: the in situ measurement is assumed to be true with zero deviation at the time of measurement.

The use and application of in situ data also require coincidental acquisition of the field sample *relative* to the acquisition of the binary image data. Restated, there is a one-to-one temporal link between the field sample and the acquired image data. Of course, this also assumes that exogenous factors such as atmospheric attenuation and sun angle do not significantly bias the binary information at the sensor level. In practice, the bias may be quite small if the images are acquired on a clear day using a nadir imaging system. Alternatively, the bias may be quite large for airborne data that are acquired during different periods of the day with time varying sun angle and atmospheric features.

The supervised approach offers many benefits for water quality analysis:

- The approach strongly weights in situ data. When these data are available and the information is properly acquired (proper sampling design), supervised classification represents a vital source of data that may be used to focus the resultant model.
- The approach is extremely stable and utilizes traditional statistical estimation as a means to separate and distinguish patterns and features.

The disadvantages of supervised classification include:

- The approach assumes in situ data are both available and of high precision. When the in situ data are of poor quality, the resultant classification is highly biased toward a class structure that is erroneous.
- The in situ information must be acquired coincidental to the binary image data.
- There is no accountability for noise or exogenous features. Atmospheric attenuation may be examined as a pre-processing step, but no accountability is provided from within the classifier.
- The approach assumes a predefined statistical method for analysis (Maximum Likelihood, Parallelepiped, etc.). No optimality conditionals are provided.

Various methods can be used in supervised classification (Richards 1999). Statistical methods are usually based on linear discriminant functions and optimal Bayesian classification using maximum likelihood (ML) algorithms. More recently, nonlinear artificial neural networks have been applied with promising results (Eurostat 1994; Cheng and Titterington 1994).

Unsupervised Classification. The unsupervised approach is advantageous when field sampling cannot be adequately conducted. This is often the case when sampling is required for large regions with limited funding, or when sampling must be conducted on a regular basis but field teams cannot be deployed to cover the target site. The approach is often the only method available when performing water quality assessments overseas. For sites with limited accessibility, image data, which can be quickly acquired at the 5- to 30-m scale, are often the only source of data for performing the required assessment.

As in the supervised case, the unsupervised approach assumes that the image data are of perfect fidelity. No consideration is provided for noisy data or exogenous features that influence the precision of the classification. Hence, image data that poor in quality will exhibit poor resultant classification accuracy in both the supervised and unsupervised case.

The unsupervised approach offers many benefits for water quality analysis:

- The approach does not depend upon in situ data. When in situ data are available and the information is properly acquired (proper sampling design), it may be used a posteriori as a technique to further refine the classification.

- The approach is extremely stable and utilizes traditional statistical estimation as a means to separate and distinguish patterns and features.

Disadvantages of unsupervised classification include:

- The approach assumes that the image data are of excellent quality. The accuracy of the classification cannot be easily established. Since in situ samples are not available, the classification represents a "best-guess" technique for identifying significant features.
- There is no accountability for noise or exogenous features. Atmospheric attenuation may be examined as a pre-processing step, but no accountability is provided from within the classifier.
- The approach assumes an a priori statistical method for analysis (Isodata). In particular, the Gaussian elements in the model do not always match the rather fragmented and discontinuous nature of environmental data. The Isodata method has limited applicability in water quality analysis since large classes (open water) are typically overrepresented in the class structure. Alternatively, subtle features (coastal patterns and near-shore features) are systematically underrepresented.

Commercial image processing applications use the Isodata algorithm as the sole analytical method. This is a straightforward application of conventional cluster analysis techniques to multivariate image data. In the area of artificial neural network theory, self-organizing nets (Kohonen 1997) offer an alternative approach, which seems closely related to the many techniques of multidimensional scaling (Cox and Cox 1994). The latter techniques focus on lower-dimensional mappings of points, and the subsequent application of standard cluster analysis.

ANALYTICAL APPROACH: This technical note focuses on non-Isodata algorithms for unsupervised classification. The approach is to develop a methodology that provides measures for the spatial dependency within an image volume. The traditional (Isodata) method assumes equal area sampling with no functional accountability for the spatial autocorrelation in the data. The next approach uses spatial measures and alternative sampling techniques to determine the separability of patterns and features. The resulting methods closely fit into the two basic characteristics for remote sensing and GIS utilization: *spatial dependence* and *spatial heterogeneity*.

Spatial dependence or *spatial autocorrelation* implies strong local associations among neighboring observations (i.e., pixels) that decline with distance (Mokken 1996). Any valid spatial analysis should incorporate this feature in its basic premises. Yet most pattern recognition analyses of remotely sensed image data usually proceed by processing each pixel's information separately and independently over the entire image, thus neglecting this basic feature. Most recent types of spatial analysis attempt to incorporate this local spatial dependence according to two approaches:

- The *neighborhood* approach, usually in the form of a spatial weight matrix $\boldsymbol{W}$ with elements w_{ij}, where for element (pixel) i, $w_{ij} = 1$ for element j when j is contiguous (a neighbor) of i, and $w_{ij} = 0$ otherwise ($w_{ii} = 0$ usually) (Anselin 1988).
- The *distance* approach, where a distance function δ_{ij} defines some distance metric between elements (pixels) i,j in the image space (Cressie 1998).

Spatial heterogeneity or *nonstationarity* is evidenced by the characteristic non-continuous variation of features across local environments in an image. This is often evidenced by areas or regions in images in which the boundaries and contours of objects are sharp or disjoint, then erratic or fuzzy (Dutilleuil and Legendre 1993). As a consequence of these two basic characteristics, spatial distributions do not fit the standard distributions of classical inferential parametric statistics and its assumptions of linearity, stationarity, and independently and identically distributed variates. Therefore, recent developments in nonparametric statistics and multivariate density estimation are used (van Kemenade, La Poutré, and Mokken 1999a).

Algorithm. The algorithm is based on the eight steps (I to VIII) shown in Figure 1. Within the first step, the raw image data are decoded from a baseline format for primary display and subsequent analysis in Steps II-VIII. The format varies by commercial vendor, but analysis is performed *after* the original binary encoded data (from the sensor system) has been "unpacked" into a digital array

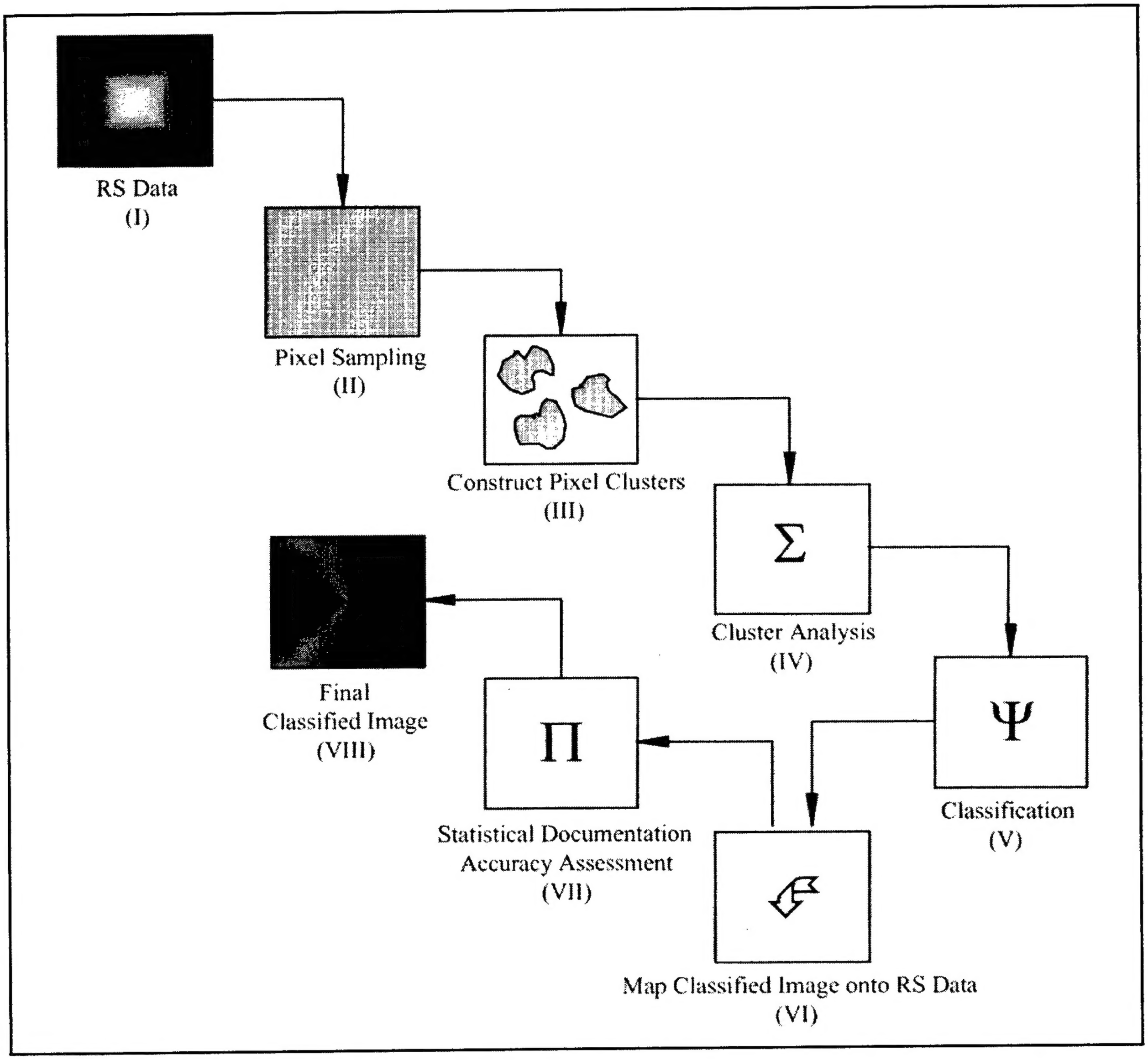


Figure 1. Central processing steps for unsupervised classification

that is conducive to image processing. Once the image is decoded, the pixel information is sampled using unbiased (traditional random) sampling or biased learning techniques (Step II). How sampling may be biased to account for pattern variation is provided in the section on biased spectral learning (see below). In Step III, preliminary clusters are formed by spectral class. The classes are organized using cluster (factor) analysis in Step IV and formal classification is performed in Step V. Step VI consists of remapping the classified information onto a new image plane of equal proportion to the original RS Data (from Step I). In this regard, the classified image in Step VI is always coregistered to the original RS data. Within Step VII, statistical documentation is provided concerning the methodology used within the unsupervised classification, and preliminary measurements are provided for determining the accuracy and precision of the final product. In Step VIII, the final unsupervised classification is displayed with "pixel-by-pixel" reporting of the class structure. The main stochastic algorithms are described in the text below.

Biased Spectral Learning. Rather than analyzing all pixels in the image, a statistically significant sample of pixel spectra is selected within Step II. Focused sampling is conducted for the following reasons:

- An adequately clustered spectral sample from the original image (in Step I) will ensure sufficient speed in performing classifications on large images. Moreover, one would like to select a set of pixel spectral values (spectral vector $\vec{s}_m$) from an image which are most informative concerning the distribution of features and class structure of that image. Not all pixels in the image fulfill this purpose, since noise and scarce (subtle) features are significantly underrepresented using a uniform sample.
- The traditional random sample (e.g. Isodata) from the image would select far too many pixels from dominant environmental strata (e.g. open water in a coastal image), and would likely miss the small nearshore features of water quality/sedimentation patterns.

Hence, the clustering and selectivity of the sampling is *adaptively* biased toward selection of pixels containing one type of ground cover ('homogeneous' or 'pure' pixels), with a low degree of noise. This is achieved by:

- Ensuring local geospatial representation *and* ensuring coverage over the image, by stratified grid sampling from (a rectangular portion of) the image.
- Selecting pixels that are spatially homogenous in the form of small locally adjacent clusters (patches), and then selecting within these patches pixels that are spectrally homogeneous (in the spectral sample *m*-space).[1]

Given a projected sample size of *N*, stratification is performed by partitioning the image within an *N* square (or rectangular) grid. Within each grid cell (stratum), a small patch of size *l* is chosen randomly. For that patch, a pixel is iteratively selected with the largest value of the local/global density ratio:

$$R^{(l,g)}\left(\vec{s}_m\right)=\frac{\hat{f}^{(l)}\left(\vec{s}_m\right)}{\hat{f}^{(g)}\left(\vec{s}_m\right)} \qquad (1)$$

[1] For the Thematic Mapper image, volume $m = 7$ spectral bands.

Within each patch, a pixel is randomly selected as a starting point and the pixel with highest density in its k_l-NN neighborhood (the median or modal point) is found. This is repeated for that point until no new points are found within the patch. This defines a modal local density $\hat{f}^{(l)}(\vec{s}_m)$ for that patch. The global spectral density distribution $\hat{f}^{(g)}(\vec{s})$ is estimated on a simple random sample from the total image and the corresponding global modal density $\hat{f}^{(g)}(\vec{s}_m)$ is analogously determined by a search of the k_g-NN neighborhood of $\vec{s}_m$. After sampling with different starting points, the spectral value of the pixel with the highest local/global density ratio *R*, typically in the range 10^2 to 10^4, is chosen as the spectral data point representing this patch in the learning sample.

Using the previous example for a coastal image that is predominately composed of open water with subtle near-shoreline features:

- The value of *R* will tend to be low for pixels that are primarily composed of open water, though obviously, a sufficient number of pixels of the open class 'water' will be selected in the learning sample, due to the dominance of open water throughout the RS image.
- The value of *R* for pixels covering the isolated water pockets and subtle near-shore features will tend to be high. Hence, there is an excellent likelihood that the algorithm will sample these features for classification analysis. Hence, the subtle features are surely represented in the classification.

The sampling methodology is designed to correct for dominant classes, in order to ensure sufficient representation in the learning sample of small, relatively scarce spectral data, such as those corresponding with littered small objects, *e.g.* oil in water, localized sedimentation patterns, tidal patterns, water density/bathymetry variations.

The Formation and Analysis of Spectral Clusters. The method of adaptive sampling selects N multispectral data points from the image volume consisting of m-dimensional spectral vectors $\mathbf{x}_i;i: l,...,N$. This sample is designed to represent the spectral distribution characteristic for the features and class structure in the image. The method of unsupervised class detection seeks to *learn* and *retrieve* the class structure from this N-sample. Spectral classes are seen as relatively dense clusters or point clouds in the spectral m-space spanned by the N-sample. Within Figure 1, the preliminary formation of spectral clusters is shown in Step III. The organization and analysis of these clusters (Step IV) is described below.

Hierarchical cluster analysis is used to locate and retrieve spectral classes. The approach is an adaptive multivariate density estimation algorithm referred to as the *k-NN estimator*. The algorithm estimates the density of the k-NN neighborhood for each sample point, as given by the m-dimensional sphere. The data point is the center of the sphere, and the (Euclidean) distance to its k^{th} nearest neighbor is the radius. This approach is intuitive for the following reasons:

- Geometrically, all points that fit within the m-dimensional sphere are organized within the cluster. Points outside the sphere cannot be reasonably positioned within the cluster since their spectral signature is not sufficient for this classification.
- The nearest neighbor estimator is numerically stable and systematically outperforms fixed kernel estimators for high-dimensional problems (van Kemenade, La Poutré, and Mokken

1999b). In this regard, the algorithm is equal to (or superior to) the traditional Isodata operation with fixed kernel representation. When the kernel happens to be equal to the fixed kernel size selected by Isodata, the methods are equivalent. In all other applications, the adaptive kernel size is numerically superior.

The k-NN density estimator for a sample point is given by:

$$\hat{f}(x) = \frac{k}{Nd_k^m x V_m} \tag{2}$$

where d_k^m is k^th^-NN distance of **x** and V_m is the volume of the unit m-dimensional ball.

Class detection is performed by means of a variant of hierarchical cluster analysis (Kaufman and Rousseeuw 1990). First, the sample points are ordered according to decreasing k-NN density. Eligible points have densities exceeding a given minimum density threshold. In high-dimensional space the 'curse of dimensionality' will hit us with very sparse densities, together with high values of k-NN distances. In this regard, a conservative value for minimum density will result in a large fraction of the learning sample not being allocated to a class. To control this problem, the algorithm uses a different dominating threshold, namely, the *fraction of points* allocated to some class.

The fraction sets a lower bound on the number of sample points to be allocated to one of the eligible clusters. The minimum density criterion is then adapted so that the total fraction of the sample allocated to some cluster at least equals that threshold. Obviously its operation implies considerable leniency concerning the minimum density criteria that follow from its application. The fraction criteria was designed to cope with the 'curse of high dimensionality' and consequently is expected to perform well for higher dimensions.[1]

Following cluster analysis, sample points are linked to spectral classes based upon proximity measures. A sample point s will be merged with a class *C* (in its proximity), when the union of its *k*-NN neighborhood with that of the nearest point *c* in that class is sufficiently dense.

The joint neighborhood of *s* and *c* is defined by the cylindrical envelope of their k-NN neighborhoods. Geometrically this may be portrayed as a cylinder with half-spherical sides. With the margin density for two clusters, c_k and c_l, (*i.e.* the density *V* of their cylindrical envelope), a criterion for their separability is defined, their *separation*, given by:

$$\mathrm{S_p} = \frac{V}{\min\{p_k, p_l\}} \tag{3}$$

where p_i denotes the maximum density in cluster c_i.

Clusters are separated when their separation *coefficient* is below a chosen threshold value, e.g. 0.50; otherwise they are merged into one class. Manipulation of the separation parameter $\mathrm{S_p}$ in Equa-

[1] The criteria will improve estimation in complex image volumes, but may bias representation for simple data sets that do not require this control.

tion 3 gives the means to vary the number of resulting classes and hence the level of detail in the classified image.

The detected clusters or classes are found as a partition of the spectral learning N-sample into |C| classes (Step V). Each class consists of at least two points, because singleton sample points are not considered as a class. However, the pixels in the image, from which these singleton spectral sample points originate, can and will be allocated to a unique class.

In summary, the cluster analysis led to *detection* and *definition* of the typical class structure of the image at the chosen levels for the system's main parameters: (a) fraction of points classified, and (b) separation of classes. Formal classification consists of assigning each pixel in the image to the best-fitting class.

Image Classification: The Allocation of Pixels to Spectral Classes. For image classification, the class structure (with corresponding linear segments and spectral distribution) is contained in the learning sample. Hence, this step requires *supervised* pixel classification.

In traditional modeling, optimal statistical techniques such as Linear Discriminant Analysis (LDA) may be applied. However, the use of LDA would require additional training of a discriminant function on the classified learning sample, which would further impede the performance of the total classification process. Therefore, a direct pixel allocation approach is required involving two alternative classification methods:

- A nonparametric nearest neighbor (1-NN) classifier. Each pixel is classified by finding the (spectrally) nearest sample point in the learning sample. The class of the nearest point is assigned to that pixel. Although a *kd-tree* is used to find the nearest neighbor, the classification time per pixel is of order log N, where N is the size of the spectral learning sample. Hence, this is the *slow* classification method.
- A projection pursuit (PP) classifier, using the first PC based on the first eigenvector in the SVD of the classes (previously described as spectral blending). For each pixel in the image, its distance to each class (i.e. line segment) is determined in terms of the coordinates for its projection (first PC) onto each class's first eigenvector. The pixel is then allocated to the nearest class. Classification time per pixel is of the order of |C|, the number of classes. Hence, this is the *fast* classification method.

The slow classification method may be used with (or without) blending. Alternatively, the fast classification method always uses blending since the distance metric for the first PC is available. Three possible classification methods may be considered:

- 1-NN classification using the slow algorithm *without* blending.
- 1-NN classification using the slow algorithm *with* blending.
- PP classification using the fast algorithm *with* blending.

The 1-NN classification using the slow algorithm *without* blending produces classification results that are nearly identical to those of the Isodata technique. The previously described methodology reduces to the direct first nearest neighbor method as used within Isodata, and only one class of information is used per pixel for mapping (portraying the image data). As a consequence, the majority of the geometric information is discarded from the image, and the final resultant classification appears as a thematic map. This is the single method used in most commercial image processing applications (Smith, Pyden, and Cole 1999).[1]

The remaining classification strategies use blending to portray the geometric information (and mixture detail) within classes. Using the blending algorithm, any pixel in the image, say with original spectral vector **x**· is represented by its projection **y** as a mixture of the endpoints of the linear first PC segment of its class. These endpoints are given by the sum of vectors **a** and **b**, where **a** indicates the first point of the first PC segment, **b** the direction vector parallel to the first eigenvector of the class cloud in the sample, and their sum the second point of the segment. Given a constant c > 0, the length of the segment, |**b**|, corresponds to 2c times the standard deviation along the first PC. Its location is chosen so that **a** + .5**b** indicates the centroid of the class cloud in the sample. Pixels with projections **y** between **a** and **b** are represented as a mixture **a** + k**b**. Pixels with projections outside that range are mapped on the nearest endpoint, so k is then set to 0 or 1. The length of the projector |**x** - **y**| is an indicator of the distance of the pixel to its class. In the image mapping within classes, blending is made to correspond with the luminance value of class color. The final classified image is stored in the form of an RGB image in GIF-format.

Statistical Documentation and Accuracy Assessment. For unsupervised classification, few measures are available for determining the accuracy and precision of a classified map. Since in situ field data are assumed to be negligible, stochastic measures must be applied to determine the *likelihood* of correct class structure. To address this requirement, the traditional Kolmogorov-Smirnov (K-S) test may be employed, since the test is nonparametric and is geometrically based upon the maximum distance between the *observed* cumulative empirical sample distribution and the *expected* cumulative sample distribution. While the test is not based upon a priori distributional assumptions (is nonparametric), it does require specific definition of an underlying function for comparison (i.e. the underlying class structure/population). The test statistic is derived as follows:

Let $X_1, X_2, \ldots, X_n$ be a random sample from a continuous distribution function F. Recall that the empirical distribution function for the sample is defined by:

$$\hat{F}_n(x) = \frac{num(X_i) \le x}{n} \tag{4}$$

where the numerator simply counts the number of samples Xi that are less than or equal to a prescribed level x and the denominator is the total number of sample points.

Under H_0: $F(x) = F_0(x)$, $x \in \Re$, we expect reasonable agreement between $F_0(x)$ and its estimator $\hat{F}_n(x)$. Conversely, suppose the alternative hypothesis H_0: $F(x) = F_0(x)$, is true for some x. Under H_0, the absolute deviation $| \hat{F}_n(x) - F_0(x) |$ should be small for all x. Hence, the following test statistic D_n is considered for classification accuracy:

[1] This option is most appropriate in the initial stages of image classification, where the primary validation of the classes is the basic issue (to be performed by investigating the class reproducing properties of various steering parameters) in connection with partial in situ data from field investigations.

$$D_n = \sup_{x \in \Re} \left| \hat{F}_n(x) - F_0(x) \right| \quad (5)$$

Thus D_n is simply the largest absolute deviation between $\hat{F}_n(x)$ and $F_0(x)$, and if D_n is small, so are the deviations.

For a 'pure' spectral class, the distribution of projections of pixels should be unimodal (binomial, or a continuous mixture of binomials). Multimodality of that distribution could indicate a 'mixed' class (and a possible source of error). For this reason, three independent K-S tests are performed against the following theoretical distributions:

- The *unimodal continuous* (Gaussian Distribution).
- The *multimodal* 4-point polygon approximation.
- The *multimodal* 8-point polygon approximation.

For the Gaussian case, the K-S statistic is calculated relative to an underlying normal distribution. The multimodal approximations are provided to examine higher-order deviations. The classification method allows the user to reference class structure against these three measures as a means to validate the classification. As in all cases where unsupervised classification is applied, high quality in situ data may be examined as a post-hoc test for classification accuracy. Given limited or nonexistent field data, the three-distribution check provides an excellent measure of class *purity*, a measure that is closely associated with classification accuracy.

EXAMPLE APPLICATION: Landsat TM scenes were acquired for two dates (8 June 1991 and 28 September 1991) coincident with intensive water quality sampling (Kennedy et al. 1994) at West Point Lake, Georgia.[1] West Point Lake is a large (surface area = 10.5 km^2; volume = 45.7 hm^3), moderately deep (maximum depth = 25 m), mainstem storage reservoir located on the Chattahoochee River 80 km downstream from the City of Atlanta in west central Georgia. Primary project purposes include hydropower, flood control, and flow regulation for navigation. The Chattahoochee River, which enters the lake approximately 56 km upstream from West Point Dam, carries heavy sediment and nutrient loads, and has a marked influence on the quality of surface waters in West Point Lake. Kennedy, Thornton, and Gunkel (1982) documented strong longitudinal gradients in nutrients, chlorophyll concentration, and water clarity, which related to advective transport, algal growth, and sedimentation. Pronounced spatial patterns in the distribution of important water quality variables related to morphologic complexity of the basin, and the differential influences of secondary tributaries have also been reported (Kennedy et al. 1994).

Recent efforts to improve West Point Lake water quality through reductions in nutrient loading from upstream sources prompted questions concerning the relative importance of local nutrient sources and led to an assessment of spatial patterns in water quality responses. Kennedy et al. (1994)

[1] An additional seven scenes were acquired for the period June 21, 1990 to April 23, 1992. Selected scenes were coincident with or were within 2 days of dates of water quality sampling efforts conducted as part of an EPA Clean Lakes Feasibility Study (personal communication, David Bayne, Auburn University, Auburn, AL). Station locations for the intensive sampling efforts were determined using GPS. Locations for the 11 stations sampled during the EPA study were approximated based on maps, buoy locations, and field observations.

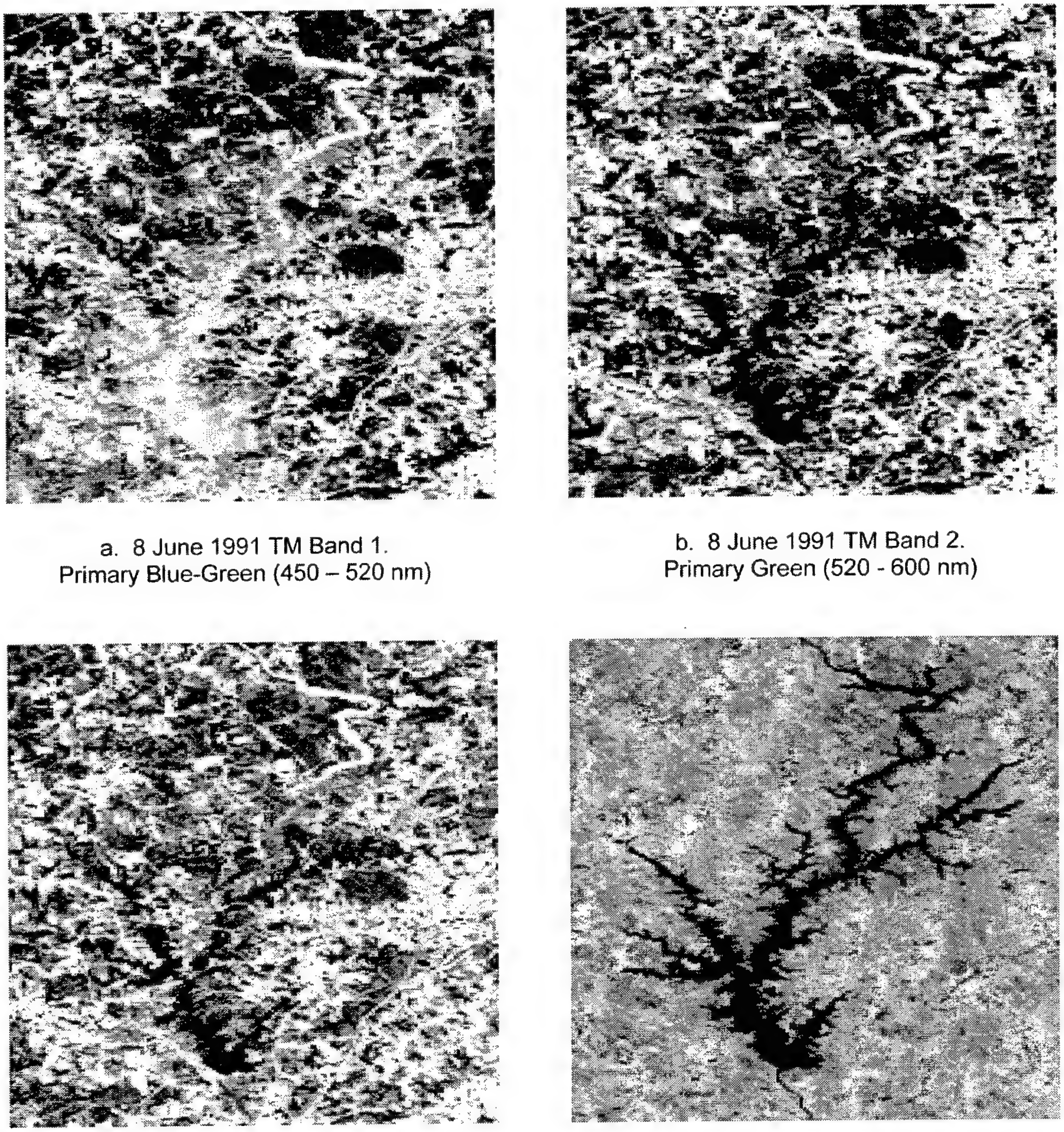
collected and analyzed surface water samples for 50-60 sampling stations located throughout the lake on dates coincident with Landsat overflights in 1991 as a means to describe spatial patterns in the distribution selected water quality variables. Variables potentially influencing reflectance values included water temperature, chlorophyll concentration, Secchi disk transparency, and turbidity.

The image layers are shown in Figures 2a-2f for the 8 June scene and in Figures 3a-3f for the 28 September scene. For this analysis, thermal band 6 was omitted to focus on 30-m multispectral data. True color composite images (3,2,1 RGB) for the two respective dates are provided in Figures 4 and 5. The true color images have been equalized for optimal separation. As a result, the image appears brighter and with greater contrast than the baseline radiometrically corrected data.

Preliminary classification results are provided in Figures 6-9. In Figures 6 and 7, raw output is given from the classification module using the following parameters: 50-percent fraction separation, 80-percent cluster separation with k-NN equal to 10 spectral neighbors. A discussion of these parameters and their logical use is provided in the "Conclusions" section. In Figures 8 and 9, the raw classification from Figures 6 and 7 has been brightness/contrast enhanced to further display the surficial features within West Point Lake. The precise correction for brightness/contrast is shown below each image.

CONCLUSIONS: In Figures 5a and 5b and Figures 6a and 6b), unsupervised classification results are provided for the two respective TM scenes. The results were generated using parameters that provided excellent separation of features as measured by the K-S statistic *D*n in Equation 5. The following justification is given for the parameter selection:

- *50-percent fraction separation.* This parameter sets the fraction of the sample points to be allocated to a cluster. The 50-percent level represents a mid-point for this allocation that is neither over-consumptive of points nor under-representative of patterns and features. The default separation is 90 percent.
- *80-percent cluster separation.* This parameter sets the cluster merging threshold. Ultimately this measure also controls the number of classes displayed in the final classification. The 80-percent level yields a large number of classes (approximately 20 classes for each of the two TM scenes). The statistical significance of the feature is not controlled by the cluster separation parameter. However, the degree of spectral blending, and the coarseness (fineness) of detail is significantly affected by this parameter. Hence, the user must iteratively test various separation levels to determine the proper level for the respective water quality application. In general, small percentages (10 to 30 percent) yield classifications with 5 to 10 spectral classes. Levels above 80 percent can produce spurious results with far too many spectral classes for adequate justification. The 80-percent level yields a nearly "unconstrained" result that generally will not limit the number of classes for visualization.
- *K-NN equal to 10 spectral neighbors.* This parameter determines the size of the square pixel window for density estimation within the hierarchical cluster analysis algorithm. The window acts as an approximate "band-pass" filter in the cluster analysis. Windows that are set too small do not adequately sample adjacent spectral levels (poor blending). Windows that are set too large consume too many samples for adequate separation of features. A default level of 10 spectral neighbors is commonly employed for excellent pattern separation.



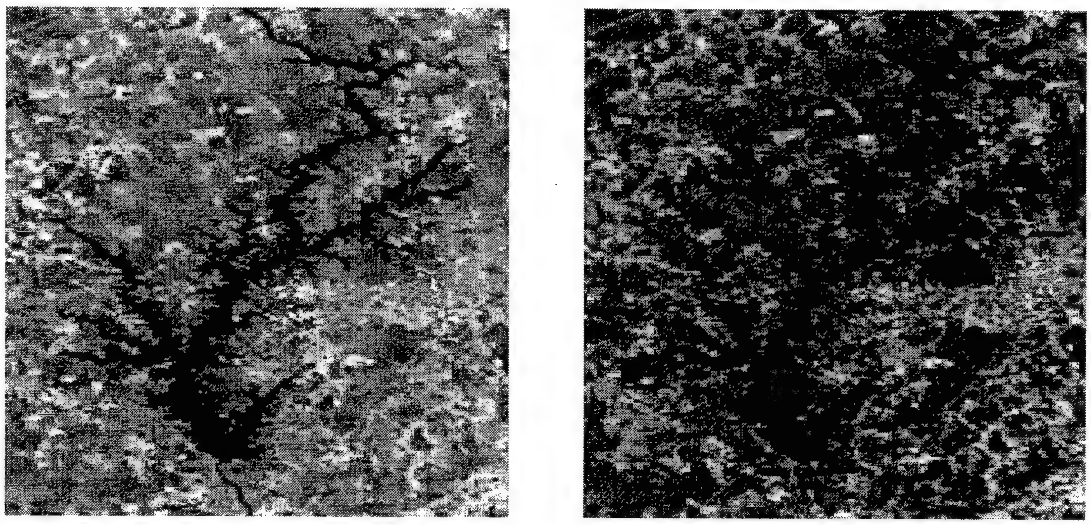
a. 8 June 1991 TM Band 1.
Primary Blue-Green (450 – 520 nm)

b. 8 June 1991 TM Band 2.
Primary Green (520 - 600 nm)

c. 8 June 1991 TM Band 3.
Primary Red (630 – 690 nm)

d. 8 June 1991 TM Band 4.
Near Infrared (760 - 790 nm)

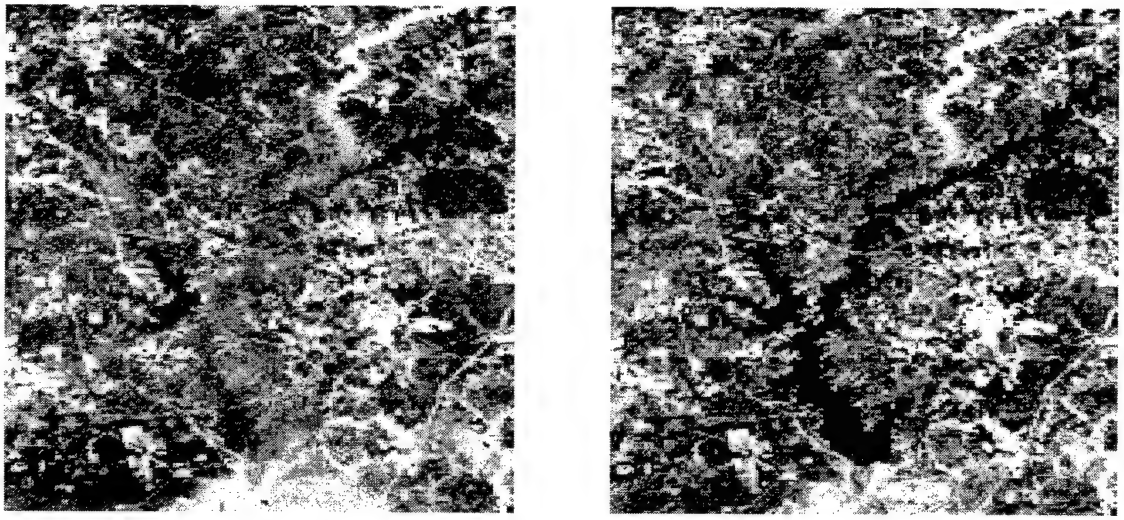
Figure 2. Image layers for 8 June (Continued)



e. 8 June 1991 TM Band 5.
Middle Infrared (1550 – 1750 nm)

f. 8 June 1991 TM Band 7.
Middle Infrared (2080 – 2350 nm)

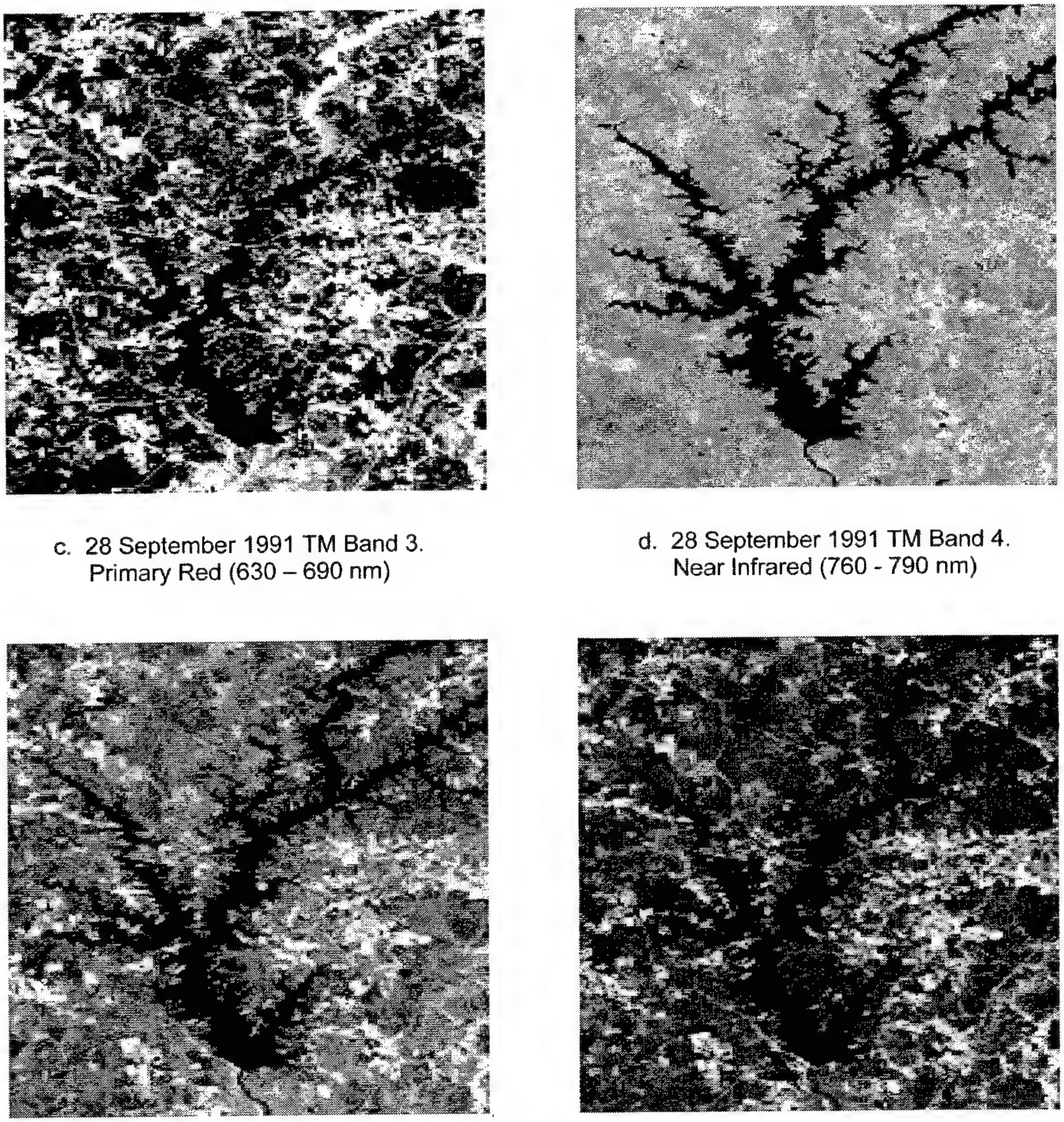
Figure 2. (Concluded)



a. 28 September 1991 TM Band 1.
Primary Blue-Green (450 – 520 nm)

b. 28 September 1991 TM Band 2.
Primary Green (520 - 600 nm)

Figure 3. Image layers for 28 September (Continued)



c. 28 September 1991 TM Band 3.
Primary Red (630 – 690 nm)

d. 28 September 1991 TM Band 4.
Near Infrared (760 - 790 nm)

e. 28 September 1991 TM Band 5.
Middle Infrared (1550 – 1750 nm)

f. 28 September 1991 TM Band 7.
Middle Infrared (2080 – 2350 nm)

Figure 3. (Concluded)

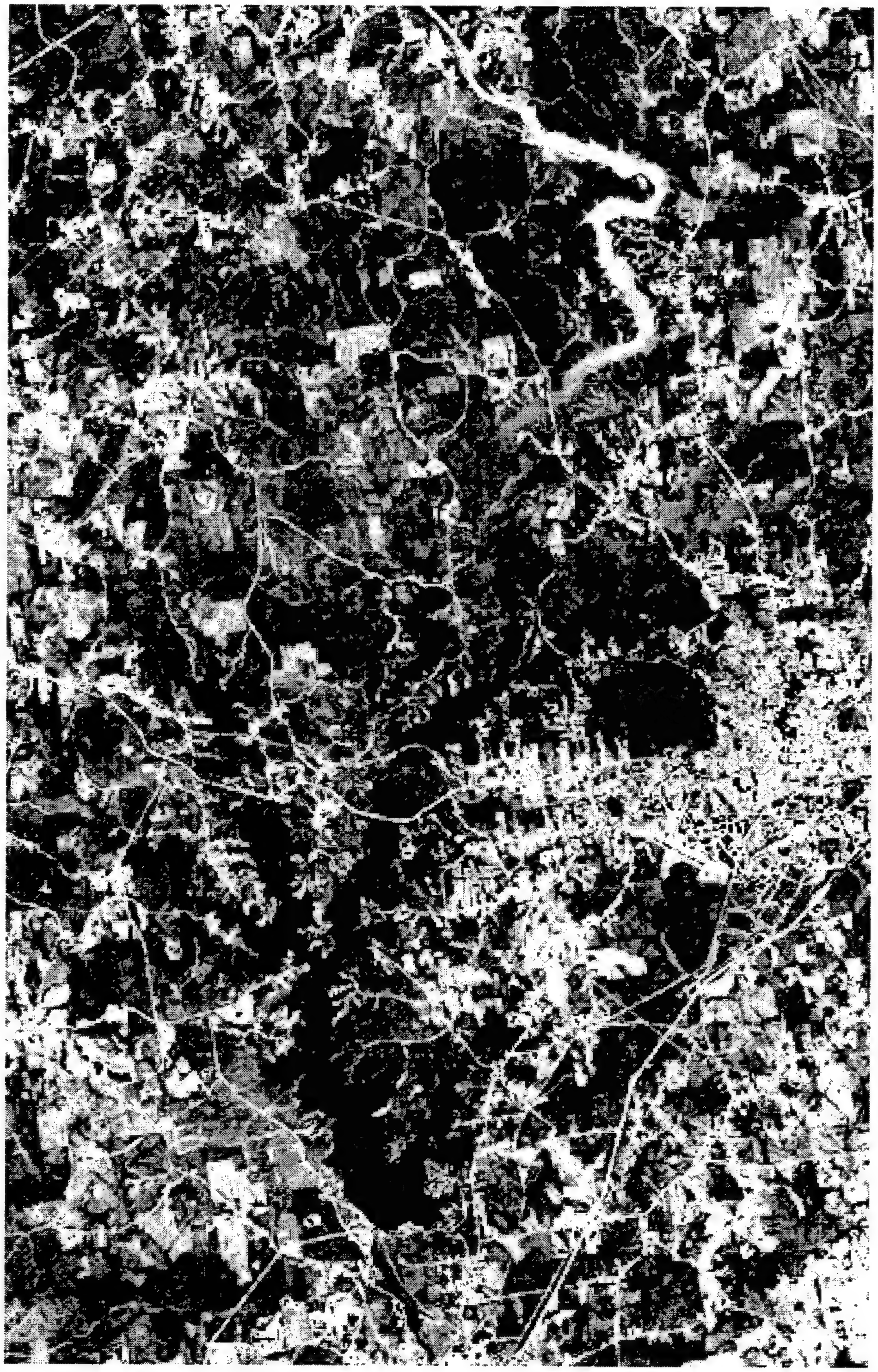


Figure 4. True color composite for West Point Lake, 8 June 1991

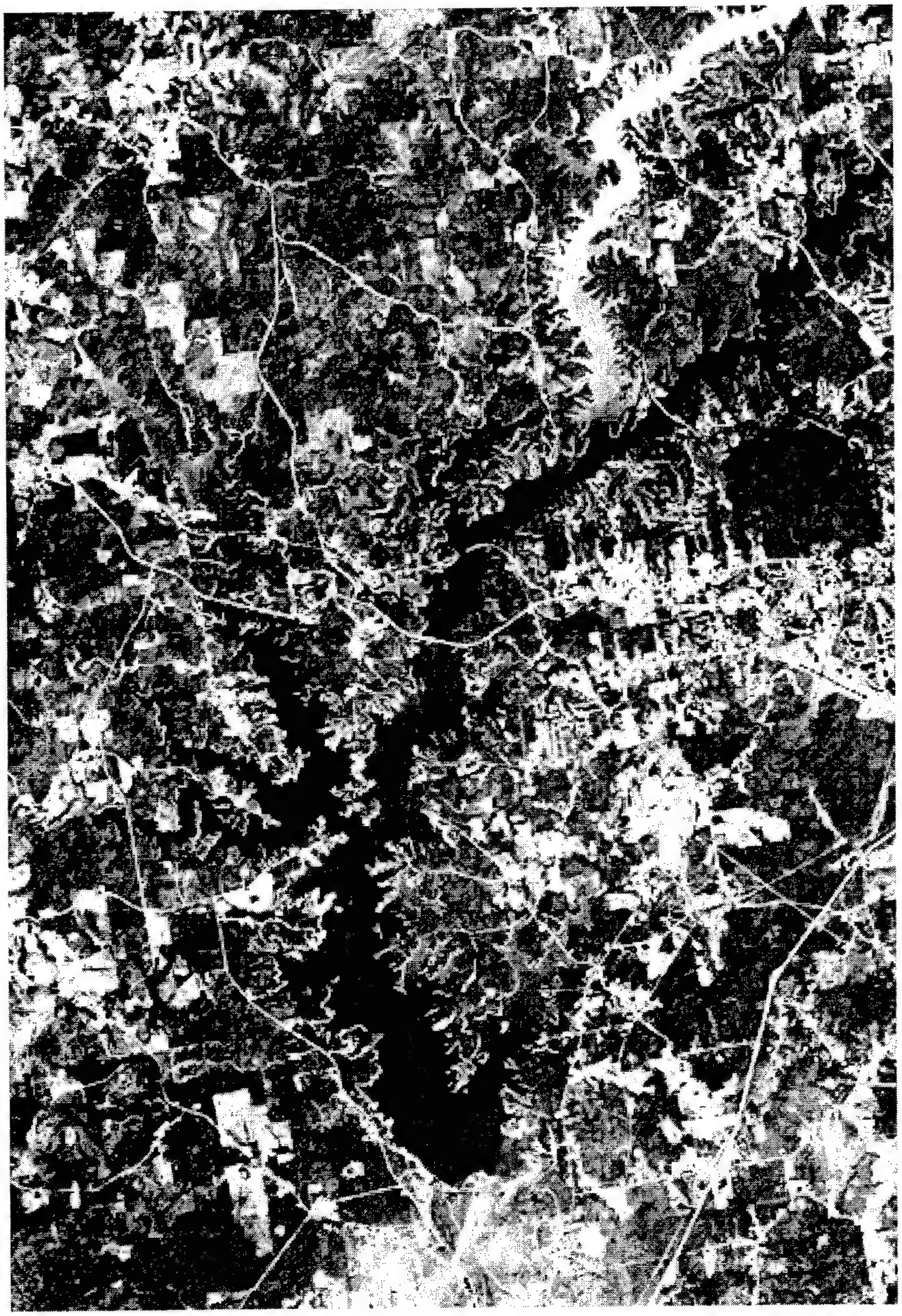


Figure 5. True color composite for West Point Lake, 28 September 1991

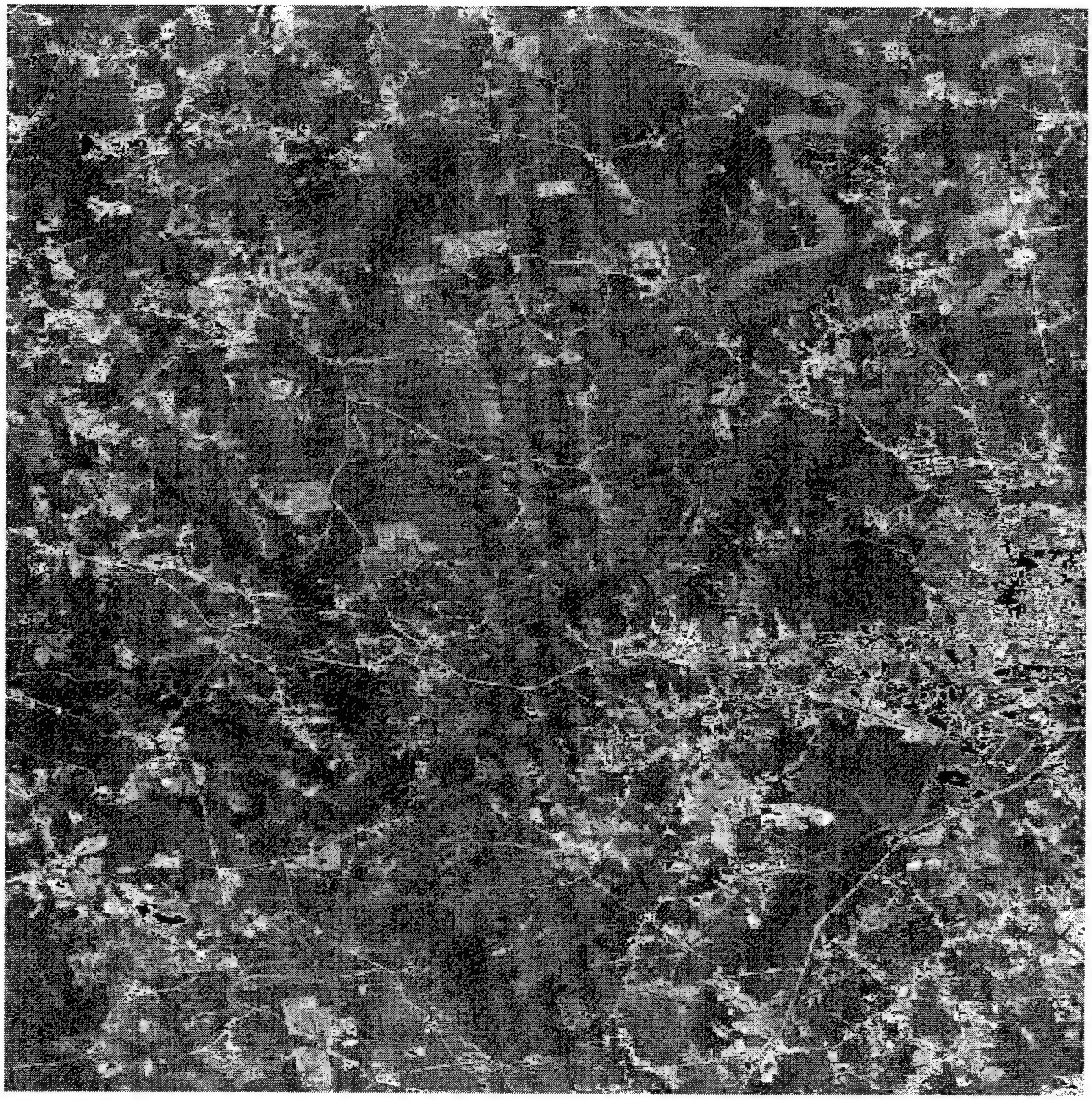


Figure 6. Primary water classification for West Point Lake, 8 June 1991

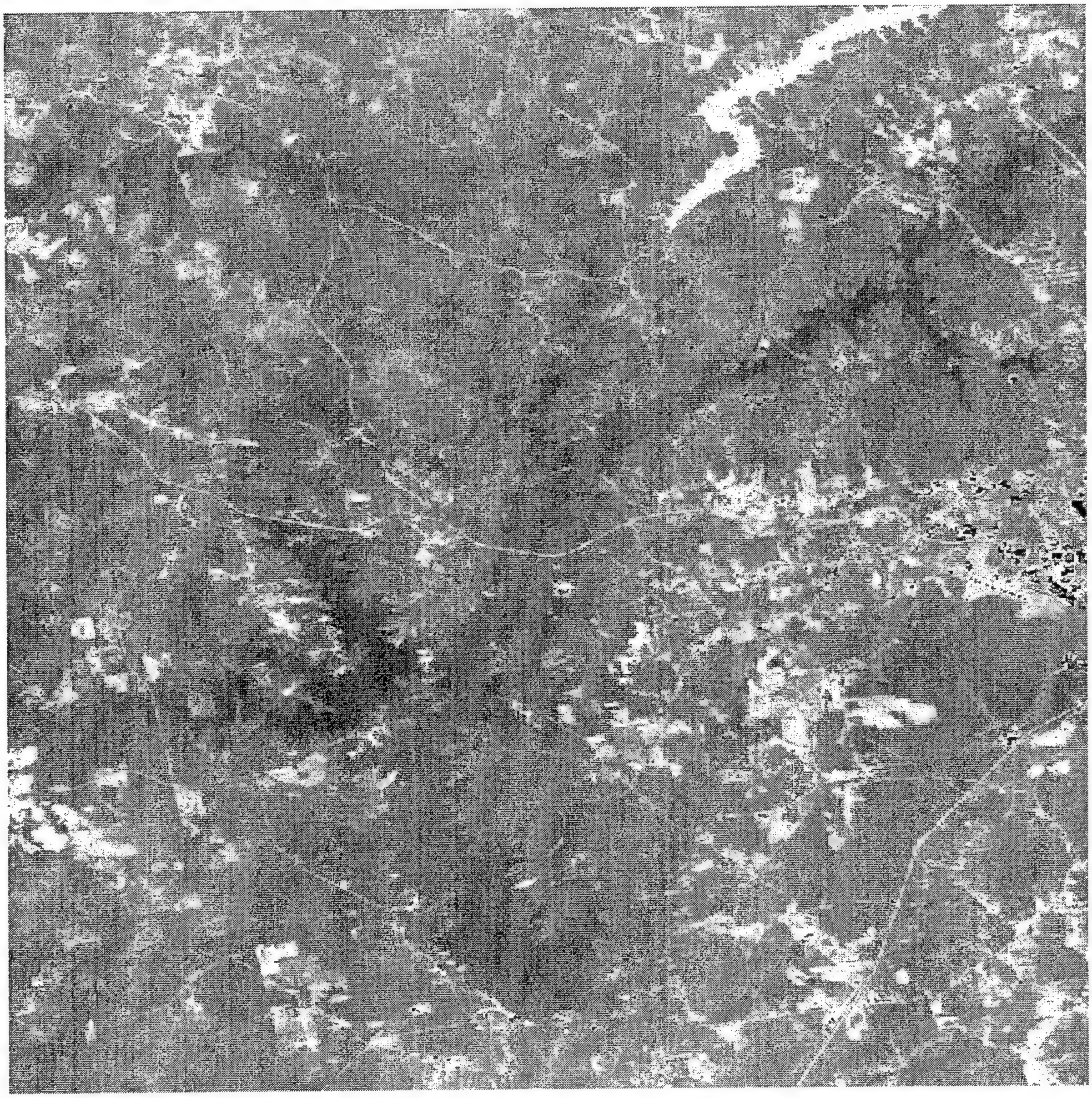


Figure 7. Primary water classification for West Point Lake, 28 September 1991

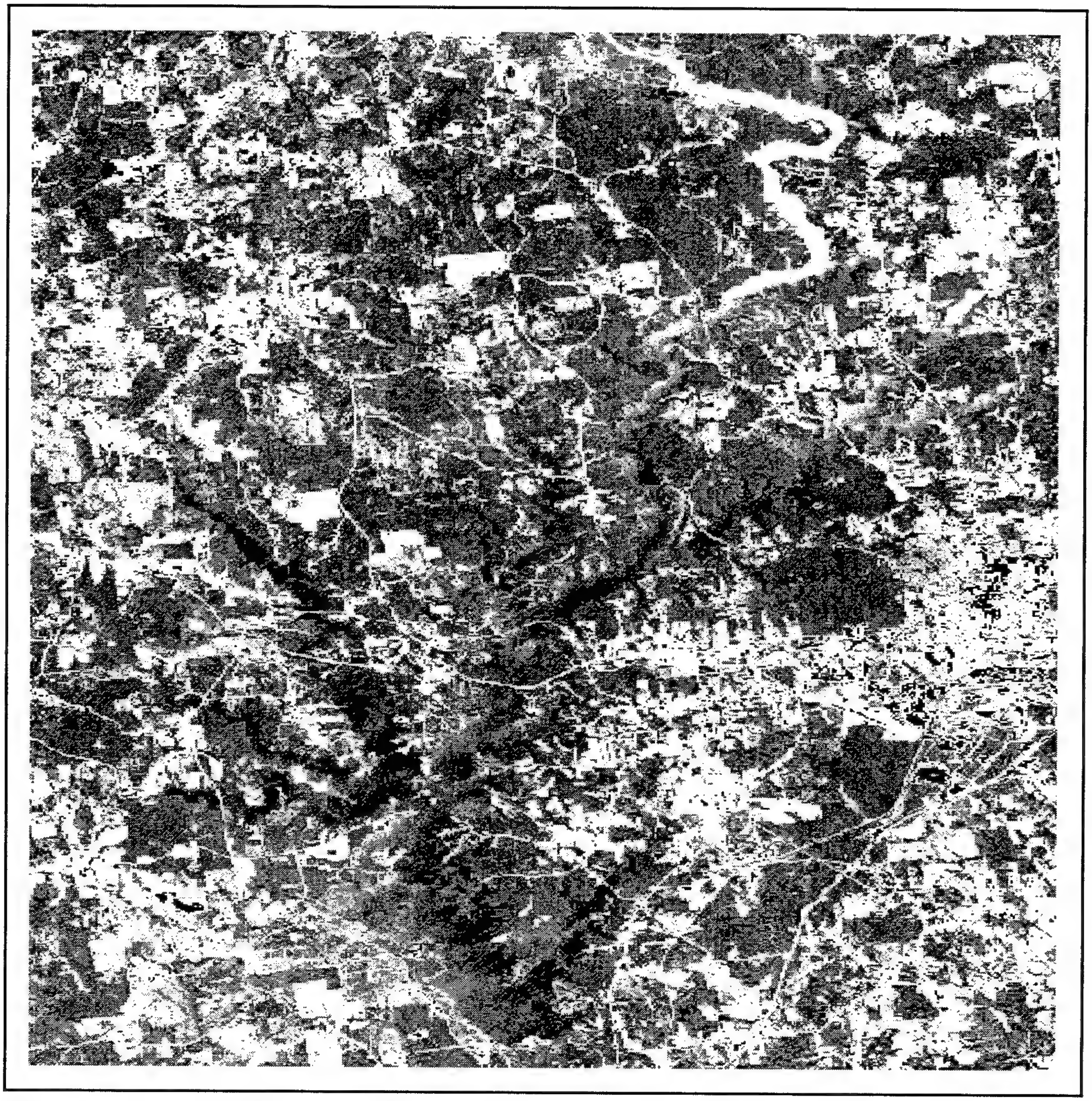


Figure 8. Water classification for West Point Lake, 8 June 1991. Brightness (-33 percent) and contrast (+78 percent) applied to enhance class structure

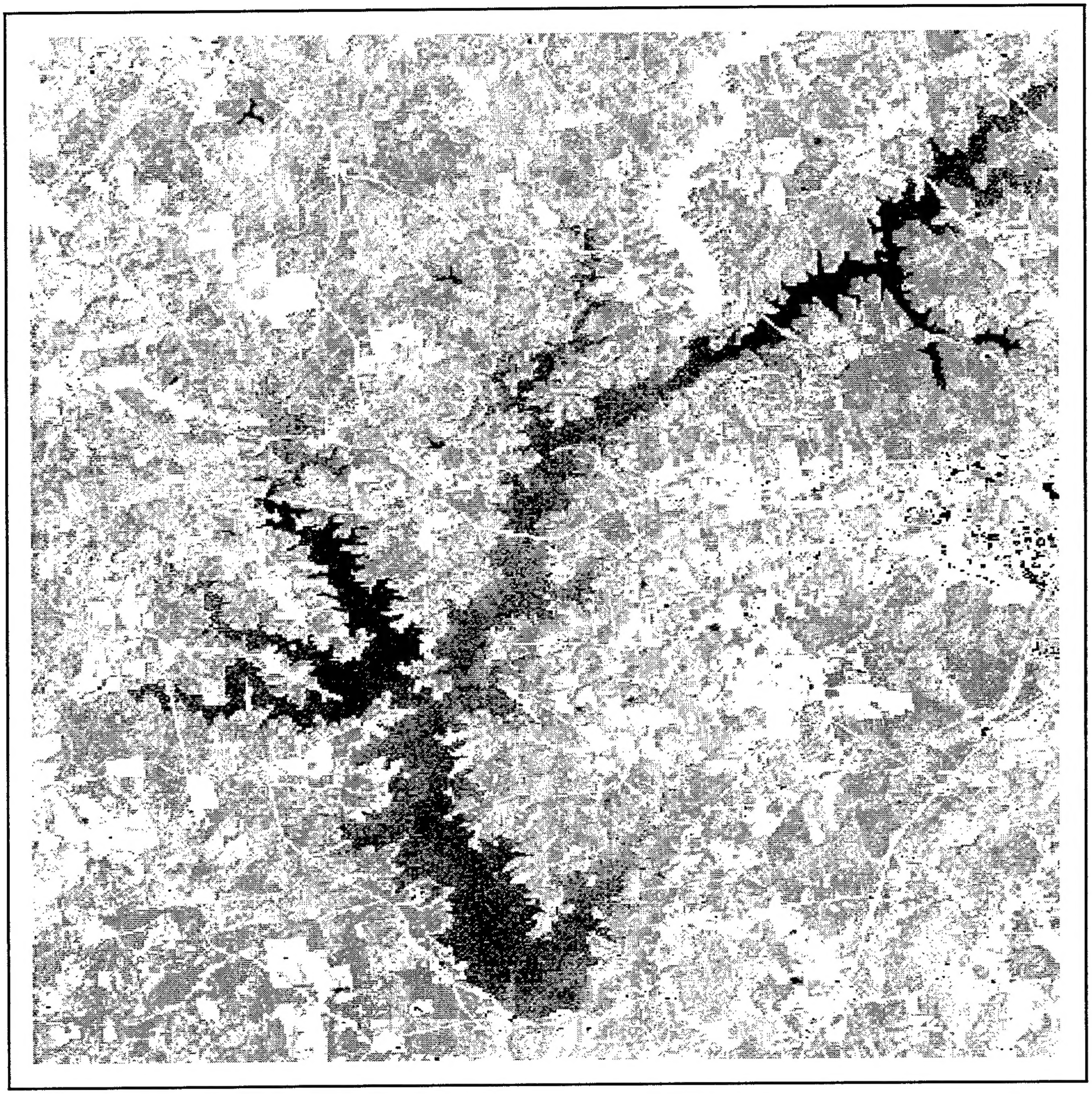


Figure 9. Water classification for West Point Lake, 28 September 1991. Brightness (-47 percent) and contrast (+85 percent) applied to enhance class structure

The classification results shown in Figures 8 and 9 are innovative for three reasons. First, the classification appears as a blended thematic map. In Figure 8, graded red intensity levels show open water. The upper channel is shown in light red primarily due to surface water reflectivity. This appears as a unique class in pink. Within the main body of the lake, three classes are shown from medium red to dark red (near-black intensity) with graded reflectivity. In Figure 9, similar features are shown along the upper reaches of the lake body. For this image, the texture is displayed using a light brown (near-yellow) shading corresponding to the primary radiometric levels in the 28 September data. As displayed, the land use characterization is minor. The primary focus is open water separation and within-class (interior) separation of surface features for West Point Lake.

Secondly, each class is generated using the focused learning sample. The classification results strongly portray all available information from small features and subtle patterns in the lake. While the class may (or may not) represent actual water quality parameters, the separation is statistically significant at the 95-percent level for all displayed classes.

Thirdly, the K-S test is shown to be significant for each spectral class within Figures 6 and 7 and Figures 8 and 9 using four spectral modes. This result indicates that the classification does not contain pure spectral classes, and additional band separation (spectral decomposition) is required. The test does not suggest the optimal strategy for separation of clusters, but does indicate that the classification will require "iterative" improvements to create optimal (unimodal) separation. The results are significantly above those generated by standard Isodata, since fewer than 10 percent of all spectral classes could be fit to four spectral modes.[1]

Assessing water quality by applying unsupervised classification is significantly improved when specific in situ samples are applied to the classification in a post hoc analysis. In this approach, the unsupervised classification is performed and the K-S test is used to test for spectral purity in each class. The in situ samples are then used as a basis to link the spectral classes to observed water quality parameters. In the case of West Point Lake, there is a reasonable correspondence between classification results for the above-described TM scenes and lake water quality data. In both scenes, regions of increasing color intensity (Figures 6-9) correspond to regions exhibiting increasing concentrations of algal pigments and decreasing levels of inorganic turbidity.

This technical note provides the detailed statistical foundations for the use and application of this methodology. This approach to RS data classification offers an opportunity to improve current Corps water quality monitoring capabilities and assess historical trends. Subsequent work will provide a test of concept using ground samples and a post hoc analysis.

[1] The classification algorithm may not yield pure spectral classes for a number of reasons:

a. The TM data are geometrically averaged to 30 m. Each pixel portrays an average intensity (by spectral band) as opposed to a pure signal (pure radiometric intensity).

b. Each TM band is of a broad spectral width. Each spectral class is composed of a large gradation of radiometric values that are stored within a single spectral band.

c. Atmospheric attenuation and other system noise. The data are not pure. The signal-to-noise ratio for this data varies by pixel position throughout the composite TM scene.

POINTS OF CONTACT: For additional information, contact Dr. Robert H. Kennedy, U.S. Army Engineer Research and Development Center, Waterways Experiment Station, (601) 634-3659, *kennedr@wes.army.mil* or the managers of the Water Quality Research Program, Dr. John Barko, (601) 634-3654, *barkoj@wes.army.mil* and Mr. Robert C. Gunkel, Jr., (601) 634-3722, *gunkelr@wes.army.mil.* This technical note should be cited as follows:

> La Potin, P., and Kennedy, R. H. (2000). "Development of unsupervised classification techniques for water quality analysis," *Water Quality Technical Notes Collection* (ERDC WQTN-AM-07), U.S. Army Engineer Research and Development Center, Vicksburg, MS. *www.wes.army.mil/el/elpubs/wqtncont.html*

REFERENCES

Anselin, L. (1988). Spatial econometrics: Methods and models. Dordrecht: Kluwer Academic.

Bruce, A. G., and Gao, H-Y. (1998). *Applied wavelet analysis with S-Plus.* New York/Berlin: Springer Verlag.

Cheng, B., and Titterington, D. M. (1994). "Neural networks: A review from a statistical perspective," *Statistical Science* 9, 2-30.

Cox, T. F., and Cox, M. A. A. (1994). *Multidimensional scaling.* Chapman & Hall, London.

Cox, R. M., Forsythe, R. D., Vaughan, G. E., and Olmstead, L. L. (1998). "Assessing water quality in Catawba River reservoirs using Landsat Thematic Mapper satellite data," *J. Lake and Res. Manage.* 14(4), 405-416.

Cressie, N. (1998). *Statistics for spatial data,* 2nd ed., Wiley, New York.

Dutilleuil, P., and Legendre, P. (1993). "Spatial heterogeneity and heteroscedasticity: An ecologocal paradigm versus a statistical concept," *Oikos* 66, 152-171.

Eurostat. (1994). "New tools for spatial analysis," *Proceedings of the Workshop, Lisbon, 18 to 20 November 1993.* Luxembourg: Office for Official Publications of the European Communities.

Feeney, R. K. (1992). "Determination of turbidity in Kentucky Lake and West Point Lake based on coincident sample and Landsat Thematic Mapper data," M.S. thesis, Murray State University, Murray, KY.

Kaufman, L., and Rousseeuw, P. J. (1990). *Finding groups in data. An introduction to cluster analysis.* Wiley, New York.

Kennedy, R. H., and Walker, W. W., Jr. (1990). "Reservoir nutrient dynamics," *Perspectives in reservoir limnology,* K. W. Thornton, ed., John Wiley and Sons, New York.

Kennedy, R. H., Hains, J. J., Ashby, S. L., Jabour, W., Naugle, B., and Speziale, B. (1994). "Limnological assessment of West Point Lake, Georgia," Technical Report EL-94-6, U.S. Army Engineer Waterways Experiment Station, Vicksburg, MS.

Kennedy, R. H., Thornton, K. W., and Ford, D. E. (1985). "Characterization of the reservoir ecosystem," *Microbial processes in reservoirs,* D. Gunnison, ed., Dr. W. Junk Publishers, Dordrecht.

Kennedy, R. H., Thornton, K. W., and Gunkel, R. C. (1982). "The establishment of water quality gradients in reservoirs," *Can. Wat. Res. J.* 7, 71-87.

Kohonen, T. (1997). *Self-organizing maps (2^{nd} ed.).* Springer-Verlag, Berlin.

Mokken, R. J. (1996). "Remote sensing, statistics and artificial neural networks - Some R&D perspectives," *Proceedings of the International Symposium on Spectral Sensing Research '95 (ISSSR), Theme: Crisis Support. 26 November - 1 December 1995, Melbourne, Victoria, Australia.* CD-ROM. Canberra: Australian Government Publishing Service.

Richards, J. A. (1999). *Remote sensing digital image analysis. An introduction. (3rd ed.).* Springer-Verlag, Berlin.

Smith, C., Pyden, N., and Cole, P. (1999). *ERDAS field guide. Edition 9.* ERDAS, Inc., Atlanta, GA.

Søballe, D. M., Kimmel, B. L., Kennedy, R. H., and Gaugush, R. F. (1992). "Reservoirs," *Biodiversity of Southeastern United States - Aquatic Communities.* T. Hackney, S. M. Adams and W. A. Martin, ed., John Wiley and Sons, Inc., New York.

van Kemenade, C. H. M., La Poutré, J. A., and Mokken, R. J. (1999a). "Density-based unsupervised classification for remote sensing." *Machine vision and advanced image processing in remote sensing.* I. Kanellopoulos, G. G. Wilkinson, and T. Moons, ed., Springer-Verlag,Berlin.

van Kemenade, C. H. M., La Poutré, J. A., and Mokken, R. J. (1999b). "Density-based unsupervised classification for multi-spectral imagery." *Spatial statistics for remote sensing.* A. Stein, F. Van der Meer, and B. Gorte, ed., Kluwer Academic Publishers, Dordrecht.

NOTE: *The contents of this technical note are not to be used for advertising, publication, or promotional purposes. Citation of trade names does not constitute an official endorsement or approval of the use of such products.*